Traffic Monitoring and Abnormality Detection Methods for Decentralized Distributed Networks

Dmytro Ageyev[a] and Tamara Radivilova[a]

[a] *Kharkiv National University of Radio and Electronics, 14 Nauka ave., Kharkiv, 61166, Ukraine*

Abstract

Internet traffic monitoring is a crucial task for the security and reliability of communication networks. This description of the traffic statistics is used to detect traffic anomalies. Modern methods of detecting attacks and other traffic anomalies in the network are not reliable enough, in particular, due to inaccurate attack moment determination, so that an attacker can easily inject errors to the operation of the system, thereby incapacitating it using DDOS attacks. To solve the problem of searching for network anomalies, a method is proposed for forming a set of informative features that formalize the normal and anomalous behavior of the system, and criteria are defined that make it possible to detect and identify various types of network anomalies. The paper discusses methods for detecting anomalies that are based on statistical approaches such as fractal traffic analysis. The issues of detection of network attacks are analyzed, which have similar statistical features, expressed in the change in mean and variance.

Keywords

Traffic abnormalities, statistical analysis, Decentralized Distributed Networks

1. Introduction

In the network security area, an intrusion is defined as a set of malicious actions against the integrity, confidentiality, and availability of information in a system or network that make it vulnerable to future attacks.

In modern conditions, one of the main processes of management of the infocommunication network is the process of managing its information security. This process must be both at the design stage of the network and at the stage of its operation and consists of a continuous analysis of the network. The effectiveness of the information security management process is significantly affected by changes in the structure, topology, and modes of the network operation. This may be due to the addition of new devices or changes to the settings of existing mechanisms, hardware failures, incorrect actions of network administrators, users, etc. When analyzing the state of the network, the main of the evaluated parameters are the probability of attack, the degree (probability) of vulnerability of network elements to information attacks. To counter threats, Intrusion Detection Systems (IDS) are deployed to detect and identify intrusion attempts into a system or network. Using a set of hardware and software resources, IDS attempt to detect intrusions by tracking data collected from a single host or network, and generate an alarm when intrusion detection is detected. IDS can be divided into different categories depending on the source of information and detection techniques.

The rapid development of computer networks and information technology raises a number of problems related to the security of network resources, which require new approaches. Currently, the issue of building intrusion detection systems is a current trend in the field of information technology. There are many papers on the topic of detecting and classifying attacks using a variety of methods, which include traditional approaches based on compliance with signature templates and adaptive models using data mining techniques.

Cybersecurity Providing in Information and Telecommunication Systems, January 28, 2021, Kyiv, Ukraine
EMAIL: dmytro.aheiev@nure.ua (A.1); tamara.radivilova@nure.ua (A.2)
ORCID: 0000-0002-2686-3854 (A.1); 0000-0001-5975-0269 (A.2)

2. Traffic Abnomality Description

Methods of detecting attacks and preventing intrusions into information systems and infocommunication networks are one of the current areas that are actively developing in the field of information security.

To do this, a number of specialized algorithms and tools are used to detect known and unknown attacks behavioral, signature methods, as well as methods to detect abnormal activity, which are particularly effective for detecting insider attacks and "zero day" attacks.

The following are used as the main classification features.

1. Source type.
2. Place of origin;
3. Method of manifestation;
4. The cause;
5. The nature of change.

To solve the problem of detecting network attacks, the most important features will be such as the source, the nature of traffic changes and the area of manifestation. The classification of network abnormalities by causes and nature of traffic changes is given in Table 1.

Table 1

Description of network traffic abnormalities

Type and cause of network abnormalities	Description	Characteristics of traffic changes
Alpha abnormality	Unusually high point-to-point traffic	Emission in the representation of traffic bytes/s, packets/s on one dominant source-destination stream. Short duration (up to 10 minutes)
DoS-, DDoS-attack	Distributed denial-of-service attack per victim	Emission in the traffic view packets/s, streams/s, from multiple sources to a single destination address.
Overload	Unusually high demand for one network resource or service	Jump in traffic on streams/s to one dominant IP address and a dominant port. Usually a short-term anomaly.
Network / port scanning	Scan the network for specific open ports or scan a single host for all ports to look for vulnerabilities	Jump in traffic on streams/s, with several packets in streams from one dominant IP address
Worm activity	A malicious program that spreads itself over a network and exploits OS vulnerabilities	Discharge in traffic without a dominant destination address, but always with one or more dominant destination ports
Point to Multipoint	Distribution of content from one server to many users	Emission in packets, bytes from the dominant source to several destinations, all to one well-known port
Disconnection	Network problems that cause a drop in traffic between one source-destination pair	The drop in traffic on packets, streams and bytes is usually to zero. Can be long-term and include all source-to-destination streams from or to a single router
Flow switching	Unusual switching of traffic flows from one inbound router to another	Drop in bytes or packets in one traffic stream and release in another. May affect multiple traffic flows.

3. Intrusion Detection Systems for Decentralized Distributed Networks

Modern networks are characterized by large volumes and speeds of information transfer. This led to inefficiency of IDS for per-packet collection and traffic analysis. Improving the IDS efficiency and productivity implements by changing methods used in IDS, moving to network traffic flow analysis using new methods based on AI.

To solve these problems, the paper offers the following IDS architecture [1], which consists of two main virtualized components: the abnormalities symptoms detection (ASD) and the network abnormalities detections (NAD). The first (ASD) is located in the network infrastructure and is its distributed component. ASD focuses on the rapid search for symptoms of abnormalities to be able to detect abnormalities in network traffic generated by user equipment and other network nodes.

On the other hand, the NAD collects timestamps and symptoms, and then the central process analyzes this data and tries to identify patterns that can be attributed to abnormal traffic. As soon as an anomaly is detected, a non-compliant message is immediately sent to the monitoring and diagnostic module.

This approach is very flexible, as it allows you to dynamically transform new virtualized resources to detect symptoms of the anomaly with increasing network traffic; and the spread, detection of symptoms, which is one of the costly processes of analysis, distributed in the network, while the detection of anomalies is centralized, which requires only symptoms in the input data.

When considering the architecture, the detection of anomalies is organized on two levels. At the lower level of the collector, the flow receives all the different flows over a given period of time, calculates the vector characteristics that the ASD module classifies as abnormal or normal. This priority classification should be done as soon as possible, even if it sacrifices accuracy for a lower call time. If an abnormal package of symptoms is suspected, which is in the window of signs, time and type of the detected anomaly, the NAD module is sent to the next level. The NAD receives several streams of symptoms from all ASDs, sorts them by time, and collects the temporal sequence of symptoms.

With this approach, it is important to learn that each ASD must support a huge amount of traffic, which is why it is extremely important to be able to select a sufficient number of threads per second, even if the detection is not quite as accurate as it may be.

4. Abnormalities Symptoms Detection

When developing a machine learning model, it is important to decide which features should be used as input data for the training algorithm. The selection of features in the formation of the feature space is a mandatory procedure both at the preparatory stage (prior to training) and at the stage of assessing the results obtained and subsequent adjustment of the training sample and/or model hyperparameters.

4.1. Statistical Characteristics which Used for Traffic Anomality Detection

To assess the current statistical characteristics of network traffic, we will use “sliding windows” of a given duration, which allow you to “view” network traffic in the “online” mode.

Statistical analysis conducted within each window can be used as a basis for constructing a space of informative features and determining the criteria for detecting network anomalies.

The analysis involves the calculation for each window of the following statistical characteristics:

- the sample mean is determined by the equation

$$\widehat{m}_i = \frac{1}{n}\sum_{j=i}^{i+n} S_j, \tag{1}$$

where S_j is the sample value of traffic intensity at the time t_j;

- the sample variance is determined by the equation

$$\delta^2{}_i = \frac{1}{n-1}\sum_{j=i}^{i+n}\left(S_j - \widehat{m}_i\right)^2; \tag{2}$$

- the asymmetry coefficient is determined by the equation

$$\gamma_{i_1} = \frac{1}{n}\frac{\sum_{j=i}^{i+n}\left(S_j - \hat{m}_i\right)^3}{\sigma_i^3}, \quad (3)$$

determining the degree of asymmetry of the probability density relative to the axis passing through its center of gravity.

The kurtosis is determined by the equation

$$\gamma_{i_2} = \frac{1}{n}\frac{\sum_{j=i}^{i+n}\left(S_j - \hat{m}_i\right)^4}{\sigma_i^4} - 3, \quad (4)$$

which shows how sharp the vertex has a probability density compared to the normal distribution. If the excess coefficient is greater than zero, then the distribution has a sharper vertex than the Gaussian distribution, if less than zero, then a flatter vertex than the normal distribution.

Antikurtosis is determined by the equation

$$\gamma'_{i_2} = \frac{1}{\sqrt{\eta}}, \quad (5)$$

where η is kurtosis parameter which determined by the equation

$$\eta = \gamma_{i_2} = \frac{\mu_4}{\sigma^4}, \quad (6)$$

where σ is standard deviation; μ_4 is sample fourth central moment.

To study the spectral properties of traffic in the absence and presence of anomalous emissions, correlation analysis is used, which includes the calculation of the correlation function, correlation coefficient and correlation interval. The calculation of the correlation function is performed according to the equation

$$R_i(k) = \frac{1}{n}\sum_{j=i}^{n+i-k}\left(s_j - \hat{m}_i\right)\left(s_{j+k} - \hat{m}_i\right), \quad (7)$$

where *k* is lag (time shift of the output row).

As the correlation coefficient $r_j(k)$ we understand the normalized value of the correlation function

$$r_j(k) = \left(\frac{R_j(k)}{R_j(0)}\right). \quad (8)$$

As the correlation interval T_{cor} we will understand the value of the argument at which the autocorrelation function for each window changes sign for the first time.

Information characteristics can be used for anomaly detection model [2]. Information characteristics, such as entropy, conditional entropy, relative entropy, information gain and information cost, are have been used. We provide the following definitions of these measures.

- Entropy is a key parameter of information theory which calculates the data collection uncertainty. For a dataset, D, the entropy is defined as

$$H(D) = \sum_{x\in D} P(x) log \frac{1}{P(x)}, \quad (9)$$

where $P(x)$ is the probability of x in D.

- Conditional entropy is the entropy of D given that Y is the entropy of the probability distribution ($P(x|y)$) as

$$H(D|Y) = \sum_{x,y\in D,Y} P(x,y) log \frac{1}{P(x|y)}, \quad (10)$$

where $P(x, y)$ is the joint probability of x and y and $P(x|y)$ the conditional probability of x given y.

- Relative entropy is the entropy between two probability distributions $p(x)$ and $q(x)$ defined over the same $x \in D$ as

$$rel\_H(p|q) = \sum_{x\in D} P(x) log \frac{p(x)}{p(q)}, \quad (11)$$

- Relative conditional entropy is the entropy between two probability distributions ($p(x|y)$ and $q(x|y)$) defined over the same $x \in D$ and $y \in Y$ as

$$rel\_cond\_H\ (p|q) = \sum_{x,y \in D,Y} P(x,y) log \frac{p(x|y)}{q(x|y)}. \quad (12)$$

- Information gain is a measure of the information gain of an attribute or feature *A* in a dataset *D* and is

$$Gain(D,A) = H(D) - \sum_{v \in values(A)} \frac{|D_v|}{|D|} H(D_v), \quad (13)$$

where values(*A*) is the set of possible values of *A* and D_v the subset of *D* where *A* has the value *v*.

Based on this knowledge, appropriate abnormalities symptoms detection models can be built. Supervised anomaly detection techniques require a training dataset followed by a test data to evaluate the performance of a model.

4.2. Analysis of the Selected Statistical Characteristics Applicability for Abnormalities Symptoms Detection

This method was tested on a set of real traffic data [3]. To do this, we used records of real network traffic, which contained attacks of various classes such as Flash-crowd, ICMP-flooding, Fraggle, Smurf, Synflooding, UDP-storm, Neptune. Experiments have shown that the impact of anomalies in the analysis window sample mean, sample variance, antikurtosis take maximum values and can be used as a symptom in intrusion detection algorithms. The kurtosis and asymmetry coefficient can also be used, but they take minimal values (Table 2)

Table 2
Traffic statistics during an attack

Attack type	$\widehat{m}_i$	$\delta^2{}_i$	γ_{i_1}	γ_{i_2}	γ'_{i_2}	*H*
Flash-crowd	max	max	min	min	max	-
ICMP-flooding	max	max	min	min	max	max
Fraggle	max	max	min	-	max	max
Smurf	-	-	min	min	max	-
Synflooding	max	max	min	min	-	max
UDP-storm	max	max	min	min	max	max
Neptune	max	-	min	min	-	max

Similarly, using records of real traffic, the analysis of information parameters of traffic was carried out. Through experiments, we observed that while the network is not under attack, the entropy values for different header fields each fall in a fairly narrow range. While the network is under attack by means of attack, these entropy values exceed these ranges.

During the experiment, not only DDoS attacks were identified, but also UDP-flood, HTTP flood, TCP SYN, Ping of Death attacks. During the analysis of the experimental results, the entropy changed for each type of attack, as shown in table 2.

Table 3
Entropy of network traffic without attacks and during attacks

w/o attack	DDoS attack	UDP-flood attack	TCP SYN attack	Ping of Death attack	HTTP flood attack
2.11	0.40	0.29	0.75	1.33	0.45
2.10	0.42	0.21	0.74	1.18	0.41
2.25	0.38	0.28	0.72	1.37	0.43
2.22	0.43	0.30	0.77	1.36	0.47

The data in Table 3 show that the entropy analysis method suitable for identifying different types of attacks.

Table 4 presents the values of the quality of attack detection by entropy analysis of protocols for different types of attacks. The proposed method has false positives. This can happen due to high traffic speed, due to limited packet header information (encrypted data), etc.

Table 4

Attack detection accuracy values for some types of attacks

Attack type	Precision	Recall	F1
DDoS	0.97	0.95	0.96
UDP-flood	0.97	0.90	0.94
TCP SYN	0.99	0.89	0.94
Ping of Death	0.96	0.93	0.95
HTTP flood	0.98	0.88	0.93

5. Conclusions

Modern information communication networks are characterized by large volumes and speeds of information transfer. Therefore, the creation of distributed intrusion detection systems, which consists of two main virtualized components: the abnormalities symptoms detection and the network abnormalities detections, makes it possible to increase the efficiency of intrusion detection.

Analysis of changes in the statistical traffic characteristics shows that during the occurrence of abnormality is observed a sharp jump of the sample mean, sample variance, entropy and antikurtosis. From the data obtained, it can be seen that the use of the above indicators in intrusion detection tasks can be used to detect abnormalities symptoms.

A method for monitoring, detecting intrusions and identifying attacks based on packet entropy analysis has been developed, which is based on the calculation of conditional entropy and statistical characteristics of packet data, which reduces intrusion detection time and identifies previously unknown attacks.

6. Acknowledgements

This work was supported in part by the National Research Foundation of Ukraine under Grant 2020.01/0351.

7. References

[1] L.F. Maimó, et al., On the performance of a deep learning-based anomaly detection system for 5G networks, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017, pp. 1-8, doi: 10.1109/UIC-ATC.2017.8397440.

[2] M. Ahmed, A. Naser Mahmood, J. Hu, A survey of network anomaly detection techniques, Journal of Network and Computer Applications, 60 (2016) 19-31. doi: 10.1016/j.jnca.2015.11.016.

[3] O.I. Sheluhin, A.S. Filinova, A.V. Vasina, Obnaruzhenie anomal'nyh vtorzhenij v komp'juternye seti statisticheskimi metodami [Detection of anomalous intrusions into computer networks by statistical methods], T-Comm - Telekommunikacii i Transport, 9(10) (2015) 42-49.